Łukasz Grabowski
Opole University, Institute of English

On translation universals in selected contemporary Polish literary translations

Summary
This pilot study attempts to examine the potential of selected corpus linguistics and computational stylistics methods in the investigation of translation universals in translational literary Polish. More specifically, the study deals with T-universals (after Chesterman 2004), which are also referred to as intralingual translation universals (Grabowski 2011), with emphasis on core patterns of lexical use, as proposed by Laviosa (1998, 2002), and the leveling-out hypothesis, as proposed by Baker (1996). To that end, the custom-designed corpora,with approximately 500,000 tokens each, of contemporary translational and non-translational literary Polish were compiled.
The results of the study reveal that on the whole translated texts are more varied lexically and have more repetitions and lower lexical variety among top-frequency words than non-translated Polish texts. On the other hand, the study shows that non-translational texts have higher lexical variety among bottom-frequency words, where usually one can find author-specific and creative vocabulary. The results of multivariate methods (Principal Components Analysis and Cluster Analysis) confirm the leveling-out hypothesis that translations are more alike as compared with native texts.

Key words:
translational texts, non-translational texts, literary Polish, translation universals, corpus-driven analysis, computational stylistics

Streszczenie
O uniwersaliach tłumaczeniowych w wybranych współczesnych polskich tłumaczeniach literackich
Niniejsze badanie o charakterze pilotażowym dotyczy wykorzystania wybranych metod badawczych językoznawstwa korpusowego i stylistyki komputerowej w analizie uniwersaliów tłumaczeniowych na materiale wybranych współczesnych polskich tłumaczeń literackich. Mówiąc ściślej, badanie dotyczy wybranych uniwersaliów typu T (za Chestermanem 2004), które nazywam uniwersaliami tłumaczeniowymi wewnątrz-językowymi (Grabowski 2011), takich jak kluczowe wzorce leksykalne (*corepatterns of lexicaluse*; Laviosa 2002) oraz hipoteza dotycząca konwergencji (*levelling-out*; Baker 1996). W celu przeprowadzenia niniejszego badania opracowano dwa specjalne korpusy badawcze (z 500 000 wyrazów tekstowych w każdym) obejmujące wybrane współczesne polskie powieści oraz wybrane współczesne tłumaczenia literackie z języka angielskiego na język polski.Wyniki badania wykazały, że jako całość teksty tłumaczone są bardziej zróżnicowane leksykalnie od tekstów nietłumaczonych, ale też cechują się większą liczbą powtórzeń i mniejszym zróżnicowaniem leksykalnym jeśli idzie o wyrazy o wysokiej frekwencji w tekście. Z drugiej strony badanie wykazało, że teksty nietłumaczone cechują się większym bogactwem leksykalnym w zakresie wyrazów o niskiej frekwencji w tekście, gdzie z reguły można znaleźć słownictwo kreatywne i odautorskie. Metody

wielowymiarowe (analiza głównych składowych, analiza skupień) potwierdziła hipotezę dotyczącą konwergencji, zgodnie z którą można zaobserwować większe podobieństwo między tekstami tłumaczonymi niż między tekstami tłumaczonymi a oryginałami napisanymi w tym samym języku.

Słowa klucze:
teksty tłumaczone, teksty nietłumaczone, polski język literacki, uniwersalia tłumaczeniowe, językoznawstwo korpusowe, stylistyka komputerowa

1. Introduction

According to Baker (1995: 233), descriptive Translation Studies (DTS) should not be limited to comparisons of source texts and their translations, but they should also be extended to comparisons of non-translated texts with translated texts, which are produced under different social, cultural and sometimes even political circumstances. In the same paper, Baker (1995: 243) puts forward an idea of universal features of translations or translation universals, which are specific textual patterns (e.g. lexical, grammatical or stylistic) typical of translated texts, irrespective of languages involved in the translation process. Further, Baker (ibid.) formulates a number of hypotheses on the differences between translational and non-translational language, e.g. that translations tend to be, among others, more explicit as regards lexis and syntax than non-translated texts, their content and form is simplified if compared with non-translated texts, and that language used in translations is more conventional and less creative than the one used in non-translated texts. As a result, translations exhibit distributions of lexical items that distinguishes them from original texts in the same language, which accounts for a symptom of specific translation strategies or tendencies, such as, among others, explicitation, simplification, normalization, sanitization and levelling-out (Kenny 2001: 53–54). In addition, characteristics of translational language are a product of constraints inherent in the translation process and do not vary across different languages and cultures (Olohan 2004: 92). Thus, it is essential to study linguistic patterns which are specific to translated texts, irrespective of source and target languages and cultures involved in the translation process (Laviosa-Braithwaite 1995: 153). Also, the analysis of translation universals with the use of corpus linguistics methods can provide further insight into the translator's presence in translation and into the style of literary translators, which is defined as "the translator's use of language, his or her individual profile of linguistic habits, compared to other translators" (Baker 2000: 245). Also, Baker adds (ibid.) that the style is a matter of patterning of linguistic features and its analysis involves describing "preferred or recurring patterns of linguistic behaviour, rather than individual or one-off instances of intervention".

Laviosa (1998: 557–570), who studied distinctive features of translational English as compared with native English (represented by samples ellicited from the Translational English Corpus (TEC) and the British National Corpus (BNC), respectively) in order to verify the existence of translation universals of simplification and explicitation, found that translational English has four core patterns of lexical use: a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words (i.e. the 200 most frequent words) over low-frequency words,

a relatively greater repetition of the most frequent words, and a smaller vocabulary (i.e. lower number of word types) frequently used. Further, Laviosa (2002: 60–62) adds that translational English texts have a lower average sentence length and a lower range of vocabulary than non-translational texts.

Consequently, a number of distinctive features of translational English in relation to native English have been uncovered. Nevertheless, if these patterns of lexical use in translational language identified by Laviosa (1998, 2002) are to be generalised as translation universals, the language pairs involved in the study must not be restricted to English (Grabowski 2012b). This observation provided motivation to undertake a project that examines features of translational Polish.

Therefore, in this pilot study, two custom-designed reference corpora of translational and non-translational literary Polish will be compared with each other in order to verify the hypotheses on core patterns of lexical use in translational texts (Laviosa 1998, 2002). Also, the study will aim to verify the so-called leveling-out hypothesis, which, according to Baker (1996: 184, quoted in Laviosa 2002: 71), provides that translational texts are more similar to one another as compared with native texts.

For the purposes of this study, the typology of translation universals [TUs] proposed by Chesterman (2004: 6–7) was applied. Chesterman (ibid.) distinguished between two types of TUs: the S-universals, which are related to translations from the source to the target language, and the T-universals, which are related to comparisons of translational and non-translational texts (i.e. target-language texts, which are not translations). In this study, which deals with comparison of translational and non-translational texts, the search for T-universals will be pursued.

2. Methodology, research material, tools and stages of the analysis

In this study, a bottom-up corpus-driven methodology was applied. In contrast to the corpus-based approach, which always works within commonly accepted frameworks of theories of language, or – in other words – is theoretically-committed (which implies prior classification of linguistic data), the reference corpora of Polish translational and non-translational texts were not adjusted to fit any predefined categories or theoretical schemata. Thus, the study questions were addressed through empirical analysis of frequency distributions of words as found in the corpus of translational (henceforth the 'PTT') and non-translational (henceforth the 'PNT') Polish texts.

When compiling these corpora, the most important criteria were representativeness and size. As regards the former, it was intended that the texts represent contemporary literary Polish. As a result, the corpus of native texts (PNT) includes a collection of literary novels published in the years 1938–1998; in a similar vein, the corpus of translations (PTT) contains Polish renditions of selected British or American novels published in Poland in the years 1954–1996.These corpora are similar in terms of size: the PNT contains 482,728 running words, whereas the PTT includes 487,760 running words.

Nevertheless, the selection and representativeness of texts to be included in the PNT and PTT is not devoid of problems of methodological character. One may pose a question concerning how reliable – the average of Polish contemporary literary fiction – are the texts included in the two corpora, or whether the reliability of these collections would change if one decided to swap some texts therein or add more texts. It is important to realize that determination of size and structure of any comparable corpus is subject to, among others, the goals of the study, the research questions, the availability of research material as well as research procedures and tools. Given so many factors, it seems that any criterion for the selection of texts may plant doubts as to whether the texts are representative of typical Polish literary fiction of the 2nd half of the 20th century. On the other hand, using statistical methods to study linguistic phenomena typical of literary language and describing texts by sequences of numbers inevitably smoothed away the effect of uniqueness or typicality attributed to individual authors (e.g. Stanisław Lem or Witold Gombrowicz). Nevertheless, the data analyzed in this study, which are not representative of all contemporary literary translations into Polish, only warrant restricted claims as to the investigated universalist hypotheses.

Detailed information concerning the make-up of the two corpora is presented in Table 1 and 2 below. The labels adjacent to titles and publication dates represent particular novels and as such they will be used in the multivariate analyses presented in section 4 of this paper.

Tab. 1. Make-up of Polish Non-Translational Corpus (PNT)

No	Author	Title, (Publication date), Symbol	Size (tokens)
1	Krystyna Siesicka	*Zapach rumianku* (1969), KS_ZR_N	30,517
2	Krystyna Siesicka	*Ludzie jak wiatr* (1970), KS_LW_N	34,566
3	Jerzy Andrzejewski	*Ciemności kryją ziemię* (1957), JA_CK_N	34,597
4	Jerzy Andrzejewski	*Bramy raju* (1960), JA_BR_N	24,867
5	Witold Gombrowicz	*Ferdydurke* (1938), WG_FD_N	71,265
6	Tadeusz Borowski	*Wybór opowiadań* (1951), TB_WO_N	58,116
7	Stanisław Lem	*Pamiętnik znaleziony w wannie* (1961), SL_PZ_N	57,576
8	Olga Tokarczuk	*Prawiek i inne czasy* (1996), OT_PI_N	55,155
9	Dorota Terakowska	*Poczwarka* (1988), DT_PO_N	68,810
10	Jerzy Pilch	*Bezpowrotnie utracona leworęczność* (1998), JP_BU_N	46,606
	TOTAL		482,728

Tab. 2. Make-up of Polish Translational Corpus (PTT)

No	Author	Title, (Publication date), Symbol	Size (tokens)
1	William S. Burroughs	*Pedał* (*Queer*) (1985), WB_P_TT	25,521
2	Jerzy Kosiński	*Malowany Ptak* (*The Painted Bird*) (1965), JK_MP_TT	59,602
3	Jerzy Kosiński	*Wystarczy Być* (*Being There*) (1971), JK_WB_TT	21,976
4	Salman Rushdie	*Wschód, Zachód* (*East, West*) (1994), SR_WZ_TT	30,183
5	Wilbur Smith	*Katanga* (*The Dark of the Sun*) (1965), WS_K_TT	67,913
6	William Wharton	*Dom na Sekwanie* (*Houseboat on Seine*) (1996), WW_DS_TT	62,065
7	William Golding	*Wieża* (*The Spire*) (1964), WG_W_TT	50,699
8	William Golding	*Władca Much* (*Lord of the Flies*) (1954), WW_WM_TT	48,997
9	Winston Groom	*Forrest Gump* (1986), WG_FG_TT	58,456
10	William Gibson	*Neuromancer* (1984) (1992), WGI_N_TT	62,348
	TOTAL		487,760

Subsequently, the quantitative corpus-driven analysis was completed with the use of WordSmith Tools 4.0 developed by Scott (2004), which is a suite of programs custom-designed for text analysis. According to Hoover (2007: 517–533), the aim of quantitative approaches to literature is to represent elements or characteristics of literary texts numerically, applying the powerful, accurate, and widely accepted methods of mathematics to measurement, classification, and analysis. Furthermore, the availability of texts in electronic format has increased the attractiveness of quantitative approaches as innovative ways of reading amounts of text that overwhelm traditional modes of reading (ibid.).

In order to obtain a measure which would quantify the degree to which the core patterns of lexical use and leveling-out universal hold, it is important to define features or characteristics of the texts to be used as a platform for comparison, or tertium comparationis. The following characteristics will be taken into consideration: (i) lexical features (lexical richness measured with the STTR and the STTR standard deviation; frequency profiles and frequency spectra); (ii) stylistic/syntactic features (mean sentence length and mean sentence length standard deviation). They will be described in greater detail throughout section 3 of this paper.

This study was broken down into two successive stages. First, the PNT and PTT were compared in terms of descriptive statistics (more specifically, the STTR, the STTR standard deviation, mean sentence length, and mean sentence length standard deviation), frequency profiles and frequency spectra. As a result, in this part of the study corpus linguistics research procedures were used. Secondly, the two corpora were

compared in terms of the distance between the 1000 most frequent words (henceforth the 'MFW') used in the texts included in PNT and PTT. In this part of the study, multivariate methods frequently used in computational stylistics and authorship attribution, namely Principal Components Analysis and Cluster Analysis, were used in order to visualize the differences between native and translational texts, and to verify whether the leveling out hypothesis – whereby there is more similarity between translations than between translations and native texts in the same language (Baker 1996: 184) – holds in the case of translational and non-translational literary Polish.

Thus, as regards validation of the hypothesis concerning the core patterns of lexical use, it is expected that (i) on the whole translational corpus (PTT) will be characterized by lower lexical richness (i.e. lower STTR); (ii) frequency profiles will show higher proportion of top-frequency words (ranked 1–200) vs. total number of words in translational texts; (iii) frequency spectra will show higher proportion of bottom-frequency (with frequencies of 1–25) words vs. total number of words.

As for validation of the leveling-out hypothesis, it is expected that (i) translational texts will show lower STTR standard deviation and lower sentence length standard deviation; (ii) Principal Components Analysis will show that distances between translational texts are lower than between native non-translational texts; (iii) Cluster Analysis will reveal the tendency of translational texts to reside in separate clusters.

3. Corpus-driven analyses

As it has been already stated above, the corpus linguistics research procedures used in the comparison of translational and non-translational Polish literary texts encompass descriptive statistics (i.e. the STTR, the STTR standard deviation, mean sentence length, and mean sentence length standard deviation) as well as frequency profiles and frequency spectra (Baroni 2009: 805–806). The results of these corpus-driven research procedures are presented in the sections 3.1, 3.2 and 3.3 below.

3.1. Descriptive statistics

Descriptive statistics describe linguistic data in quantitative terms, which is commonly accepted in corpus linguistics as basic indicators of style and lexical richness (Olohan 2004, 78–81). Hence, it provides a holistic view of the two corpora of Polish texts (PNT and PTT), whose characteristics are presented in Table 3.

Tab. 3. Descriptive statistics for PNT (native texts) and PTT (translations)

Statistics	PNT	PTT
Number of running words	482,728	487,760
Number of word tokens (used for a wordlist)	478,504	487,362
Number of word types	70,722	65,412
Type/token ratio (TTR) (% or *x* per 1000 tokens)	**14.77**	13.42

Statistics	PNT	PTT
Standardized TTR (STTR) (in % or *x* types per 1000 tokens)	60.40	**62.24**
STTR (mean across texts)	59.74	**62.19**
STTR standard deviation (STTRstd)	**39.75**	36.04
STTRstd (mean across texts)	**38.20**	36.03
Mean word length (in characters)	5.32	**5.36**
Number of sentences	42,936	50,183
Mean sentence length (in tokens)	**11.14**	9.71
Mean sentence length standard deviation (in tokens)	23.37	7.12

Table 3 shows that the two corpora are not of the same size, which is due to the fact that they contain full-texts of novels instead of their samples. Thus, the fact that PTT contains almost 10,000 more word tokens than PNT has to be taken into consideration while comparing lexical richness with the use of either the TTR or the STTR. In general, these two indicators provide brief information as to the complexity/simplicity or specificity/generality of a particular text or corpus–lower TTR or STTR translates into narrow range of vocabulary, or lower lexical richness in a text or corpus. However, the TTR is highly sensitive to differences in size of texts or corpora. As a rule, shorter texts have higher TTR value, and longer texts have lower TTR value (which is due, among others, to continuous repetition of grammatical words). As a result, although the TTR shows that non-translational texts are lexically richer, one must turn to the STTR, which is calculated on consecutive 1,000-token-long fragments of text and then averaged out, for more reliable information (Scott 2007: 157). Consequently, the STTR shows that on average translations (PTT) have 622 word types per 1,000 tokens, whereas in PNT there are only 604 word types per 1,000 tokens. The STTR calculated separately for each text in PTT and PNT, and then averaged out produces similar results (622 and 597 in PTT and PNT, respectively). Summing up, this perfunctory measure of lexical richness shows that on the whole translational texts are more complex and specific lexically and have fewer lexical repetitions as compared with native texts. As a result, the core pattern of lexical use – translational texts having lower range of vocabulary – was invalidated.

On the other hand, higher values of the STTR standard deviation in the PNT show that some text fragments in non-translational texts are lexically richer than the ones in the PTT. Also, the lower value of the STTR standard deviation is paramount to lower lexical dispersion (or variability) in translational texts, which means that they are more homogenous and similar to each other. This finding confirms the leveling-out hypothesis as regards lexical richness in translational texts.

As regards the mean sentence length, there are differences between the PNT and PTT (11.14 and 9.71 words, respectively). The former score is similar to the one revealed in the study conducted by Ruszkowski (2004: 34), where the mean sentence

length for Polish prose was found to be 11.90 words. The discrepancy between the PNT and PTT enables one to formulate a number of hypotheses. Firstly, longer sentences in the native texts can show that their style is more explicit and precise as compared with concise and terse sentences in the PTT. Secondly, longer sentences may indicate that they are more lexically varied and that there is higher information load therein. Moreover, the mean sentence length standard deviation shows that the PNT has less uniform distribution of sentences, which translates into higher syntactic variability (the high mean sentence length standard deviation of 23.37 signals that among the sentences in the PNTone may identify conspicuously long ones). Overall, one may hypothesize that the style of the translations is more uniform in terms of length of sentences (mean sentence length standard deviation of 7.12), which confirms the leveling-out hypothesis with respect to syntactic variability. On the other hand, the style of typical literary texts (non-translational) is more varied, i.e. one is bound to find there both conspicuously short and long sentences.

To sum up the analysis and interpretation of descriptive statistics, one may hypothesize that on the whole the translational texts are, on average, lexically richer than non-translational texts, which is surprising with the view of the hypothesis of lexical simplification in translation. Longer sentences in non-translational texts and their less uniform distribution there may show that they are more explicit and precise than the ones found in the translations. On the other hand, it may also mean that there is a tendency to use simple and concise sentences in the translations, which is in line with the hypothesis of syntactic simplification in translation. Nevertheless, it seems that further and more detailed qualitative research is essential to bring to life concrete illustrations of the above differences and find the rationale underlying them, e.g. are they due to the authors' style, or translators' style, or the interference from English (as it is a source language for all texts collected in the PTT) and thus the direction of translation?

3.2. Frequency profiles

In order to determine whether translational texts (PTT) or non-translational Polish texts (PNT) have more repetitions and lower lexical variety among top-frequency words (i.e. the words ranked 1–200 on a frequency list (Laviosa 1998), a frequency profile proposed by Baroni (2009: 805–806) was used. As a rule, the frequency profile is obtained by a replacement of words in a frequency list (which was completed with the use of WordSmith Tools 4.0) with their frequency-based ranks, by assigning rank 1 to the most frequent word, rank 2 to the second frequent word, rank 3 to the third frequent word etc. It enables one to determine which frequency-based ranks (r) of words (tokens) have a particular frequency (f). However, in this study a typical frequency profile was modified in that frequency information was substituted with information on cumulative percentage of the total word count (%cW) corresponding to frequency-based ranks. The results are presented in Table 4 below.

Tab. 4. Frequency profiles for top-frequency word types in PNT and PTT

Non-translational texts (PNT)		Translational texts (PTT)	
Rank	%cW	Rank	%cW
1	3.17	1	3.11
10	17.80	10	18.03
20	22.68	20	22.64
30	25.45	30	25.43
40	27.57	40	27.41
50	29.28	50	29.02
60	30.79	60	30.43
70	32.11	70	31.72
80	33.23	80	32.83
90	34.26	90	33.82
100	35.16	100	34.68
200	**41.26**	**200**	**40.45**
518	**50.01**	546	**50.01**

The data show that translational texts (PTT) feature no significantly different distribution of top-frequency words. As a rule, the higher the share of top-frequency words in the total word count (%cW), the less lexically varied and more repetitious a text or corpus. As a result, Table 4 shows that translational texts have fewer repetitions and higher lexical variety among top-frequency words (the words ranked 1–546 account for 50% of the total word count in PTT, while in the case of PNT there are only 518 words that reach this threshold).

In order to verify whether these differences are statistically significant or random in the case of words ranked 1–200, Dunning's (1993) log-likelihood test at the probability value p = 0.05 was used. As a matter of fact, the word types ranked 1–200 encompass 197,483 word tokens in PNT and 197,300 word tokens in the PTT. The comparison of the two values conducted with the use of the said test yielded the LL-score of 36.67 for words ranked 1–200), which means that the differences between the PNT and PTT are statistically significant.[1] Thus, one of the core patterns of lexical use in translational texts put forward by Laviosa (1998) that translations have a rela-

[1] As a 2 × 2 contingency table was used to compare the frequencies of words ranked 1–100 in the PNT and PTT, a degree of freedom (d.f.) equals 1, which means that according to chi-square distribution table (which is the same for the log-likelihood test) one can reject the null hypothesis whereby there is no statistically significant difference between observed frequencies in the PNT and PTT only if the critical value of the log likelihood exceeds 3.84 (at p = 0.05).

tively higher proportion of high-frequency words (i.e. the 200 most frequent words) over low-frequency words (i.e. words ranked 201 and lower) than native texts, or a relatively greater repetition of the most frequent words, was invalidated in the case of translational Polish literary texts as attested in the PTT.

3.3. Frequency spectra

According to Baroni (2009: 806), frequency spectra enable one to determine how many word types (w) in a frequency list have a particular frequency [w (f)]. As creative or author-specific vocabulary usually occurs in a text with low frequencies, frequency spectra can be used to study lexical variety and number of repetitions among bottom-frequency words (i.e. words with frequencies 1–25). As a rule, a text is more varied lexically if the proportion of bottom-frequency words in the total word count (%W) is higher. For the purposes of this study, the number of word types (w) corresponding to particular frequency (f) in the frequency spectra was substituted with information on the cumulative percentage of the total word count (%cW) corresponding to word types with frequencies 1–25. The results are presented in Table 5.

Tab. 5. Frequency spectra for PNT and PTT

PNT		**PTT**	
w (f)	%cW	w (f)	%cW
1–25	37.20%	1-25	36.29%

The data show that non-translational texts have higher lexical variety among bottom-frequency words, which account for 37.20% of the total word count. More specifically, there are 178,008 word tokens with frequencies 1–25 in the PNT and 176,883 in the PTT. In order to verify whether these differences are statistically significant or random, Dunning's (1993) log-likelihood test at the probability value p = 0.05 was used. The comparison of the numbers of words with frequencies 1–25 with the use of the said test yielded the LL-score of 54.05, which means that the difference between PNT and PTT is statistically significant. Overall, it means that Polish literary translations have lower lexical variety among bottom-frequency words than native texts, which confirms one of the core patterns of lexical use as postulated by Laviosa (1998). Such a T-universal can also mean that translators tend to use more conventional target language expressions while rendering creative vocabulary attested in the original novels. Nevertheless, further qualitative studies are required to validate this hypothesis.

4. Multivariate analyses

The analyses above focused on the data sets with only two vectors presented in a tabular form, i.e. specific observations (numbers) in the rows, with the vectors (column variables) specifying different properties of the observations (e.g. 11.14 being a mean sentence length in the PNT). This part of the study focuses on data sets with more

than two vectors. More specifically, the aim of multivariate analyses presented below is to verify if the translational texts (PTT) are similar or different than individual literary texts contained in the Polish Non-Translational Corpus (PNT) in terms of the differences (i.e. distance) between frequencies and distributions of the 1,000 most frequent words[2] (MFW) in each of the texts.

As a matter of fact, multivariate analyses have been frequently used methods in computational stylistics(see: Burrows 2004: 323–347, Craig 2004: 273–288 for an overview of research studies), which is a quantitatively rigorous study of linguistic patterns in texts that are linked to the processes of writing and reading, i.e. to style in the wider sense of the word (Craig 2004: 273). The popular variants of multivariate analysis are Cluster Analysis, Correspondence Analysis and Principal Components Analysis. Also, according to Eder and Rybicki (2011: 308), in the last decade stylometrists have commenced to develop custom-designed statistical methods, such as Burrow's Delta, Zeta and Iota (Burrows 2002, 2006) and their modifications by other scholars (Argamon 2008, Craig and Kinney 2009, Hoover 2004a, 2004b, all quoted in Eder and Rybicki 2011: 308). It is therefore worthwhile investigating whether the above research methods typical of computational stylistics, computational stylometry and authorship attribution can be utilized in the study of translation universals, and the leveling-out hypothesis, in particular.

Thus, the aim of the following research procedures is to measure and visualize the distance between translational and non-translational texts (i.e. the shorter the distance, the more similar the texts) in order to verify whether the leveling out hypothesis– whereby there is more similarity between translations than between translations and native texts in the same language (Baker 1996: 184) – holds in the case of translational and non-translational literary Polish.

The distance between texts will be measured with the use of a Delta method proposed by Burrows (2002), which is a simple measure of difference between two texts frequently used in authorship attribution and computational stylistics (Hoover 2004, Rybicki and Eder 2009, Popescu and Dinu 2009). In simple terms, Delta measures distance between z-scores of the pair of words which occur in two texts with a given frequency (Burrows 2002, quoted in Popescu and Dinu 2009: 351). The distance measured by Delta method will be presented in the form of two-dimensional graphsgenerated throughout Principal Components Analysis and Cluster Analysis.

As a result, this part of thestudy deals with data sets with more than two vectors, i.e. 1000 different word types (MFW), their frequencies (f) and 20 texts, which follows that the number of vectors (and correspondingly, dimensions) is high. The data sets like this are referred to as multivariate data and they are typically brought together in matrices (Baayen 2008: 127).

[2] Studying frequencies and distributions of word types and word tokens in texts, Burrows (1987) and Baayen (2001), among others, questioned the idea that word-tokens appear randomly in texts. Using multivariate analyses (more specifically, multidimensional methods), they showed that there is a powerful 'force-field' attached to each occurrence of a word.

Overall, the analyses and visualizations were completed with the help of the Rscript[3] developed by Eder and Rybicki[4] (2011). More specifically, the visualization was completed through the application of multivariate statistical techniques used to find structure in data through groupings of observations (Baayen 2008: 127), namely Principal Components Analysis (henceforth 'PCA') and Cluster Analysis (henceforth 'CA'), which are presented in greater detail below.

4.1. Principal Components Analysis

PCA is a statistical technique used to identify and express patterns in data of high dimension in such a way as to visualize their similarities and differences. Since patterns in data of high dimension can be hard to find and visualize, PCA is a powerful tool for the analysis (Smith 2002). As it was already mentioned, the data set under investigation encompasses 1,000 different word types (MFW) and their frequencies in 20 texts, which follows that the number of dimensions is high. Thus, it is expected that the PCA will enable one to squish such a high number of dimensions into two and visualize the distances between the texts found in the PNT and PTT in two dimensional space. As a rule, PCA aims to derive a relatively small number of variables to convey as much information in the observed variables as possible (Leech et al. 2005: 88). In other words, it enables one to use fewer variables (i.e. principal components) to provide the same information that one would obtain from a larger set of variables (ibid.). It is therefore hoped that among 1,000 MFW the PCA will identify a few words, which can be labelled 'style discriminators', and that these words will provide as much information about the differences between the two sets of texts as one would obtain from analysis of frequencies and distributions of 1,000 or 10,000 MFW.

The PCA was completed with the use of the aforementioned R script developed by Eder and Rybicki (2011). As a result,the script generated the following (Fig. 1) two-dimensional graph, which presents the distances (in terms of 1,000 MFW meas-

[3] The advantages of using R for data analysis are thoroughly described by Baayen (2008, viii) and Gries (2009). R is an interactive programming environment, an open-source implementation of the object-oriented source language for statistical analysis (exploratory data analysis). Once one has mastered its grammar and acquired its basic vocabulary, it is possible to easily complete advanced data analysis specifying selected statistical models, no matter which type of model is being fitted. As R has outstanding graphical facilities, it provides a number of visualization methods, which generally provide far more insight into the data than longish lists of statistics. For more information, see: http://cran.r-project.org.

[4] The R script was developed by Maciej Eder and Jan Rybicki during ESU 2009 and ESU 2010 "Culture and Technology" in Leipzig. In this study, the version 0.3.7 of the script was used (M. Eder, personal communication, March 8, 2011), and it is downloadable, including newer versions, from the website: https://sites.google.com/site/computationalstylistics/scripts. The script was presented to the public in June 2011 at the Digital Humanities 2011 conference held in Stanford, CA. In a nutshell, it enables one to perform various analyses in computational stylistics as it supports a number of nearest neighbor classification methods used in stylometry. According to the authors (Eder and Rybicki 2011), the script is a free, open-source (GPL licensed), and cross-platform software, which is supplemented with a graphic user interface, which makes it easily adjustable to a wide-range of research purposes.

ured with Delta method) between individual texts found in the PNT (these texts are marked with the letter _N at the end of the labels[5]) and PTT (these texts are marked with the letters _TT at the end of the labels).

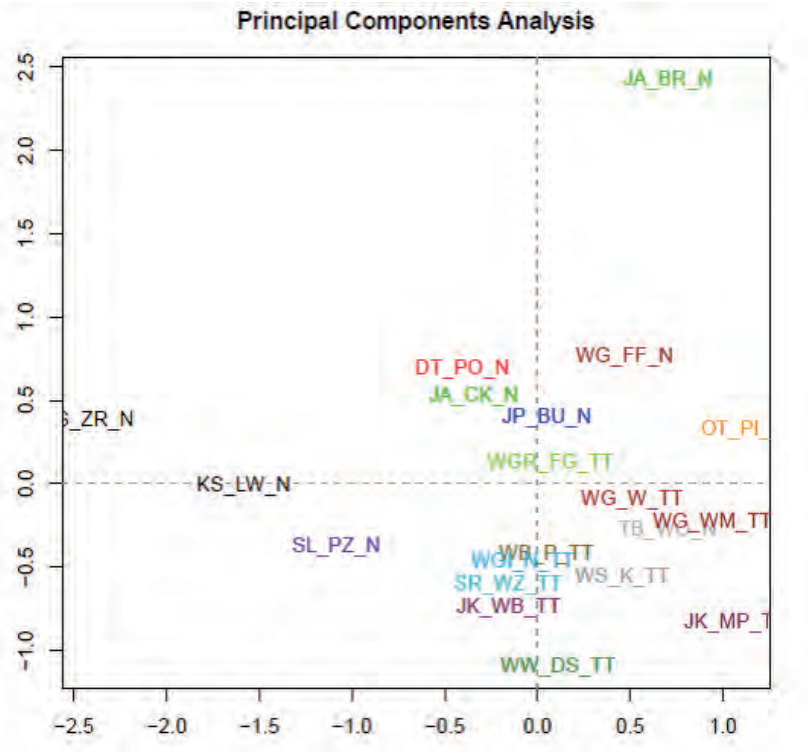

Fig. 1. Principal Components Analysis (PCA) of the texts from the PNT and PTT

Fig. 1 reveals that all translational texts (_TT) are placed, roughly, in the rightmost bottom quarter of the graph, suggesting more uniform distribution within this set of texts. On the other hand, the non-translational texts (_N) are scattered around all four quarters of the graph, which makes the distance between them and translated texts even more pronounced. This observation, namely more uniform distribution and more similarity between translated texts in terms of frequencies and distributions of 1,000 MFW, corroborates the aforementioned leveling out hypothesis that there is more similarity between translations than between translations and native texts in the same language.

Nevertheless, there remains a question concerning specific words which to the highest degree impacted the positions of texts in Fig. 1. Thus, the PCA biplot was completed in order to identify the principal components, i.e. specific words of the biggest discriminating strength. These are presented in Fig. 2 below.

[5] The labels representing particular texts found in the PNT and PTT are presented in Table 1 and Table 2 above.

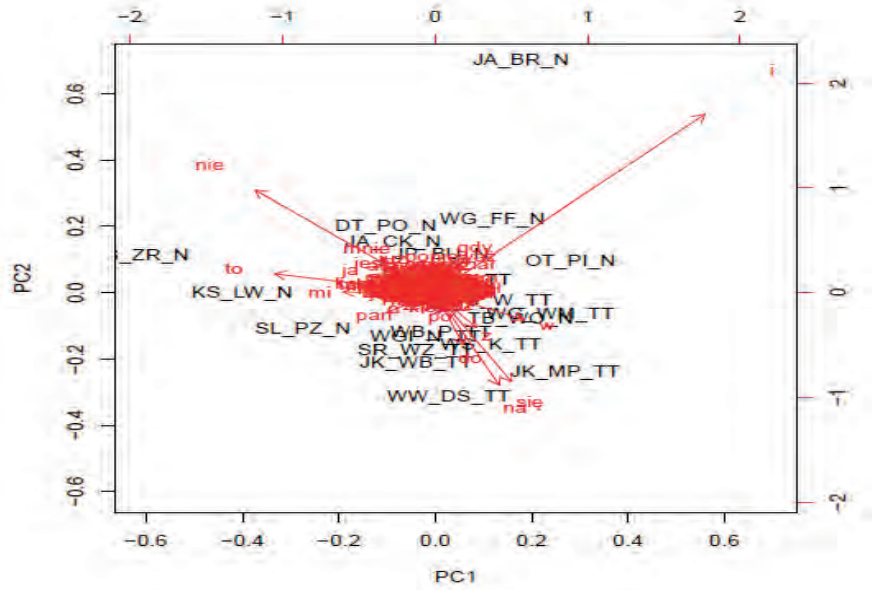

Fig. 2. PCA biplot

The biplot presented above reveals that the words *i* (conjunction 'and'), *nie* (negative particle 'no' or 'not'), *na* (preposition 'on'), *się* (reflexive pronoun '-self'), *to* (demonstrative pronoun 'this') and *w* (preposition 'in') are the most important stylistic discriminators (principal components) and that the remaining words – contained in the grey spot in the centre – have weak discriminating strength. In other words, it means that the arrangement of the non-translational and translational texts in Fig. 1 is largely determined by six the most discriminating words specified above, which is not paramount to the actual differences between these texts. As a matter of fact, authorial or translator's style is certainly something more than the frequency of a few words. It is due to the inherent feature of PCA, which is designed for extracting the most important information, i.e. the principal components, from the data under investigation and ignoring the vast majority of observations (Leech et al. 2005: 88). In practice, in a study like this one, PCA treats the most frequent variables (words) as the most important ones because these words are very frequent and their frequencies in a set of texts are significantly spread (M. Eder, personal communication, March 6, 2011). As a result, the leveling-out hypothesis is corroborated only on the basis of data presented in Fig. 2. It seems that the PCA procedure should be repeated and conducted on contents (lexical) words only, or with the top-frequency words filtered out altogether to render more comprehensive and reliable results and further validate or invalidate the leveling-out hypothesis.

4.2. Cluster Analysis

According to Tan et al. (2006: 487), Cluster Analysis (CA) is a multivariate statistical technique used to divide the data into conceptually meaningful groups of objects sharing common characteristics. CA groups data objects (e.g. texts) based only on

information found in the data that describes the objects and their relationships. In this study, the information used in CA is a distance between 1,000 MFW in the texts measured on the basis of Delta method (likewise Rybicki and Eder 2009, Popescu and Dinu 2009) . In other words, this technique enables one to automatically find similar objects and display them in a tree-like format (Baayen 2008: 148) in clusters, which are also known as dendrograms. As a rule, the greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better and more distinct the clustering (Tan et al. 2006: 490). Thus, CA provides an abstraction from individual data (i.e. texts) to clusters in which these data objects reside. It is therefore regarded as a form of classification in that it creates a labeling of objects (i.e. texts) with class (cluster) labels derived only from the data which describes the said objects. As a result, such a classification will constitute unsupervised classification (ibid.: 491).

Eventually, the use of Cluster Analysis produced the following (Fig. 3) graph, which displays the distance, calculated on the basis of the Delta method (Burrows 2002), between all individual texts contained in the PNT and PTT in the form of clusters grouping classes of similar texts.

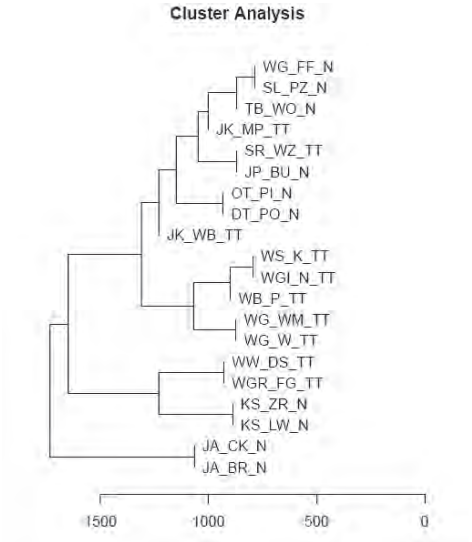

Fig. 3. Cluster Analysis (CA) of all individual texts contained in PNT and PTT

A dendrogram presented in Fig. 3, which is a hierarchical tree plot, provides further evidence that translated texts (PTT) differ from non-translated texts (PNT) in terms of frequency and distribution of 1,000 MFW. With the exception of one text – a Polish translation of Salman Rushdie's *East, West* – translational and non-translational texts reside in separate clusters. As a result, CA presented above validates the leveling-out hypothesis that there is more similarity between literary translations than between literary translations and native texts in Polish. Thus, the degree of similarity (or association) is maximal between non-translated texts included in the PNT, which are all grouped into one cluster, with smaller sub-clusters amalgamated into one. Again,

one is made to conclude that translated texts (PTT) differ from typical native literary texts included in the PNT.

Nevertheless, illuminating as they may seem to be, both PCA and CA reveal only one thing, namely that translated texts differ from non-translated texts, yet they do not explain why these differences exist. These observations bring one to conclusions and suggestions for future work, which are presented in the following section.

5. Conclusions and future work

The aim of this pilot study was to verify translation universals hypotheses, namely the core patterns of lexical use and the leveling-out treated as T-universals (Chesterman 2004), with the use of selected corpus linguistics and computational stylometry research procedures applied to custom-designed corpora of translational and non-translational literary Polish.

As regards verification of the hypothesis concerning the core patterns of lexical use, (i) it was revealed that on the whole translational texts are characterized by higher lexical richness (higher value of the STTR), which invalidated this core pattern of lexical use; (ii) comparison of frequency profiles showed that translational texts have fewer repetitions and higher lexical variety among top-frequency words (i.e. the ones ranked 1-200 on a frequency list), which invalidated the core pattern of lexical use whereby translations have more repetitions and lower lexical variety among top-frequency words; (iii) comparison of frequency spectra revealed that translations have lower lexical variety among bottom-frequency words (i.e. the ones with frequencies 1–25) than native texts, which confirmed one of the core patterns of lexical use. As a result, one in three core patterns of lexical use was validated in this study.

As for verification of the leveling-out translation universal, (i) it was revealed that translational texts are more homogenous and more uniform as regards lexical richness and sentence length (lower STTR standard deviation and mean sentence length standard deviation as compared with non-translational texts), which confirmed the leveling-out hypothesis; (ii) multivariate analyses, such as the Principal Components Analysis and Cluster Analysisconfirmed the leveling-out hypothesis that translations are more alike as compared with native texts in terms of the distance–measured with the use of Delta (Burrows 2002) – between frequencies and distribution of the 1,000 most frequently used words.

Although one can put forward a number of hypotheses on differences between translational and non-translational Polish literary texts on the basis of the results presented above, it seems that further qualitative research should be conducted to bring to life concrete illustrations of both typical and anomalous cases glossed over in this study. It is vital since it is still unknown what factors (and to what an extent) impact the basic stylometric indicators presented throughout this study. Notwithstanding some attempts at providing preliminary answers to that question (Rybicki 2009), the very impact of a source language and target language, direction of translation, genre-specific characteristics, text type, register characteristics, translator's idiolect, author's idiolect, translator's and author's ideologies, source language culture, target language culture

onto basic stylometric indicators and, more generally, onto the scope and character of language universals still remain a debatable issue and account for rather unexplored research area, particularly in the case of Polish language material (Grabowski 2012a).

Therefore, more quantitative and qualitative research on translation universals in Polish (in line with studies by Scarpa 2006, Corpas Pastor et al. 2008, Xiao 2010), including examination of other T- and S-universals (i.e. explicitation, simplification, normalization, sanitization and leveling out) should be conducted in the future. To that end, it seems that compilation of larger corpus of translational Polish – including texts representing different genres and produced by different translators – should be the step in the right direction.

Also, in order to eliminate possible source-language bias (in this study, all the texts in the PTT are Polish translations of novels originally written in English), the corpus of translational texts should include translations from different languages (Baker 1995: 234).

Finally, as demonstrated by the results of the Principal Components Analysis and Cluster Analysis, and by measuring distance between translational and non-translational texts with the use of a Delta method developed by Burrows (2002), the untapped potential of using computational stylometry, computational stylistics and authorship attribution methods in research on translation universals should be further explored in the future. In particular, it is suggested that the leveling-out hypothesis in translational literary texts be further verified with the use of PCA and CA, and other neighbor classification methods, such as Burrows' Delta, Theta and Iota, Argamon's Delta, Eder's Delta, Manhattan Distance, Canberra Distance or Euclidean Distance.

References

Baayen Harald (2001): *Word frequency distributions*. — Dordrecht: Kluwer.

Baayen Harald (2008): *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. — Cambridge: Cambridge University Press.

Baker Mona (1993): Corpus linguistics and Translation Studies: Implications and applications — [In:] Mona Baker, Gill Francis, Elena Toginini-Bonelli (eds.): *Text and Technology: In Honour of John Sinclair*; Amsterdam: John Benjamins, 233–250.

Baker Mona (1995): Corpora in translation studies: An overview and some suggestions for future research. — *Target* 7 (2): 223–243.

Baker Mona (1996): Corpus-based Translation Studies: The Challenges that Lie Ahead. — [In:] Harold Somers (ed.): *Terminology, LSP and Translation: Studies in Language Engineering. In Honour of Juan C. Sager;* Amsterdam: John Benjamins, *175–186.*

Baker Mona (2000): Towards a Methodology for Investigating the Style of a Literary Translator — *Target* 12 (2): 241–266.

Baroni Marco (2009): Distributions in text. — [In:] Anke Lüdeling, Merja Kytö (eds.): *Corpus linguistics: An international handbook Volume 2*; Berlin–New York: Walter de Gruyter, 803–821.

Burrows John (1987): *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* — Oxford: Clarendon Press.

Burrows John (2002): "Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship. — *Literary and Linguistic Computing* 17 (3): 267–287.

Burrows John (2004): Textual Analysis. — [In:] Schreibman, Simens, Unsworth, 323–347.

Chesterman Andrew (2004): Beyond the particular. — [In:] Anna Mauranen, Pekka Kuyamaki (eds.): *Translation Universals: Do they exist?*; Amsterdam: John Benjamins, 33–49.

Corpas Pastor Gloria, Mitkov Ruslan, Afzal Naveed, Pekar Victor (2008): "Translation universals: do they exist? A corpus-based NLP study of convergence and simplification". — [In:] *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)* (available at: http://www.amtaweb.org/papers/2.05_Corpas.pdf (accessed April 2012)).

Craig Hugh (2004): Stylistic Analysis and Authorship Studies". — [In:] Schreibman, Simens, Unsworth, 273–288.

Dunning Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. — *Computational Linguistics* 19 (1): 61–74.

Eder Maciej, Rybicki Jan (2011): Stylometry with R". — *Digital Humanities 2011: Conference Abstracts*; Stanford CA: Stanford University, 308–311.

Grabowski Łukasz (2011): Korpusy dwu- i wielojęzyczne w służbie tłumacza, leksykografa i badacza: poszukiwanie ekwiwalentów przekładowych w świetle hipotez dotyczących istnienia uniwersaliów tłumaczeniowych. — [In:] Wojciech Chlebda (ed.): *Na tropach translatów. W poszukiwaniu odpowiedników przekładowych*; Opole: Wydawnictwo Uniwersytetu Opolskiego, 89–112.

Grabowski Łukasz (2012a): *A Corpus-Driven Study of Translational and Non-Translational Texts: the Case of Nabokov's 'Lolita'*. — Opole: Wydawnictwo Uniwersytetu Opolskiego.

Grabowski Łukasz (2012b): Between Stability and Variability. A Corpus-Driven Study of Translation Universals: the Case of Polish Translations of 'Lolita'. — [In:] Liliana Piasecka, Ewa Piechurska-Kuciel (eds.): *Variability and Stability in Foreign and Second Language Learning Contexts: Volume 1*; Newcastle upon Tyne: Cambridge Scholars Publishing, 2–24.

Gries Stefan (2009): *Quantitative Corpus Linguistics with R. A Practical Introduction*. — New York–London: Routledge.

Hoover David (2004): Testing Burrows's "Delta". — *Literary and Linguistic Computing* 19 (4): 453–475.

Kenny Dorothy (2001): *Lexis and Creativity in Translation*. — London: Routledge.

Laviosa-Braithwaite Sarah (1995): Comparable corpora: towards a corpus linguistic methodology for the empirical study of translation. — [In:] Marcel Thelen, Barbara Lewandowska-Tomaszczyk (eds.): *Translation and Meaning Part 3*; Maastricht: UPM,153–163.

Laviosa Sarah (1998): Core patterns of lexical use in a comparable corpus of English narrative prose. — *Meta* 43 (4): 557–570.

Laviosa Sarah (2002):*Corpus-based Translation Studies. Theory, Findings and Applications*. — Amsterdam–New York: Rodopi.

Leech Nancy, Caplovitz Barret Karen, Morgan George (2005): *SPSS for intermediate statistics: use and interpretation*. — London: Routledge.

McEnery Tony, Xiao Richard,Tono Yukio (2006): *Corpus-based Language Studies: An Advanced Resource Book*. — London–New York: Routledge.

Olohan, Maeve (2004): *Introducing Corpora in Translation Studies*. — London–New York: Routledge.

Popescu Maria-Lidia, Dinu Michaela (2009): Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis". — [In:] *Proceedings of the International Conference RANLP-2009*; Borovets: Association for Computational Linguistics, 349–354 (available at: http://www.aclweb.org/anthology/R09-1063 (accessed April 2012)).

Ruszkowski Marek (2004): *Statystyka w badaniach stylistyczno-składniowych*. — Kielce: Wydawnictwo Akademii Świętokrzyskiej.

Rybicki Jan, Eder Maciej (2009): PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff? — *Digital Humanities 2009: Conference Abstracts*; College Park, MD: University of Maryland, 242–244.

Rybicki Jan (2009): Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read*?* — *Digital Humanities 2009: Conference Abstracts*; College Park, MD: University of Maryland, , 245–247.

Scarpa Federica (2006): Corpus-based Quality-Assessment of Specialist Translation: A Study Using Parallel and Comparable Corpora in English and Italian. — [In:] Maurizio Gotti, Susan Sarcevic (eds.): *Insights into specialized translation*; Bern: Peter Lang Verlag, 155–172.

Schreibman Susan, Simens Ray, Unsworth John, eds. (2004): *A Companion to Digital Humanities*. — Oxford: Blackwell.

Scott Mike (2004): WordSmith Tools version 4. — Liverpool: Lexical Analysis Software.
Scott Mike (2007): WordSmith Tools Help. — Liverpool: Lexical Analysis Software.
Smith Lindsay (2002): *A Tutorial on Principal Components Analysis*. *Cornell University* (available at: http://users.ecs.soton.ac.uk/hbr03r/pa037042.pdf (accessed March 2010)).
Tan Pang-Ning, Steinbach Michael, Kumar Vipin (2006): *Introduction to Data Mining*. — Addison: Prentice Hall.
Xiao Richard (2010): How different is translated Chinese from native Chinese? A corpus-based study of translation universals. — *International Journal of Corpus Linguistics* 15 (1): 5–35.